Pivot-Based Bilingual Dictionary Creation for Low-Resource Languages(Abstract_要旨)

AUTHOR(S):

Mairidan, Wushouer

CITATION:

Mairidan, Wushouer. Pivot-Based Bilingual Dictionary Creation for Low-Resource Languages. 京都大学, 2015, 博士(情報学)

ISSUE DATE:

2015-03-23

URL:

https://doi.org/10.14989/doctor.k19117

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	買日旦 吾守爾 （Mairidan Wushouer）
論文題目	Pivot-Based Bilingual Dictionary Creation for Low-Resource Languages （低資源言語のためのピボット型対訳辞書生成）		

（論文内容の要旨）

The goal of this thesis is to support the construction of new language resources for low-resource languages. To this end, it details a new approach to creating bilingual dictionaries of intra-family languages based on the pivot-based technique. The thesis consists of seven chapters.

Chapter 1 outlines the thesis, including the research objective, approaches and issues.

Chapter 2 describes the background of the thesis, the research problem and existing studies on bilingual dictionary creation. To this end, the word sense ambiguity problem, a key problem to be solved in automatic bilingual dictionary creation, and its existing solutions are described in detail. For the sake of comprehensibility, the existing studies are classified into various groups according to the type of the language resource used to create a new bilingual dictionary, such as parallel corpora, comparable corpora and bilingual dictionaries. In addition, the creation of bilingual dictionary of intra-family languages is also discussed from the viewpoint of the etymological closeness of the languages.

In Chapter 3, a heuristic framework of bilingual dictionary induction is proposed.
Large scale language resources are very useful for automatic acquisition of high quality bilingual dictionaries, but such resources are inaccessible for low-resource languages. Therefore, it is important to make best use of the limited available language resources. To this regard, a framework is designed to induce a new bilingual dictionary of an intra-family language pair from the incorporation of various types of language resources as well as human effort. To realize this, the heuristics is defined as a function to measure the relativeness of a cross lingual word pair based on certain criteria, and a list of heuristics are extracted from one or a group of language resources, which are then incorporated by a mathematical model to estimate overall semantic relativeness. The key insight of the framework are as follows: (1) the ability of creating heuristics from the structure of bilingual dictionaries by using a high-resource language as a pivot between two intra-family languages, (2) incorporating predefined heuristics to estimate semantic relativeness of the cross-lingual word pairs, and (3) an iterative induction mechanism that can produce the new bilingual dictionaries in different quality range. To evaluate the framework, this chapter also details an experiment in which a bilingual dictionary of Uyghur and Kazakh languages, the members of Turkic language family, was created by using basic heuristics which were extracted from spelling similarity of these two languages and their existing bilingual dictionaries with Chinese, a member of Sino-Tibetan language family. As a result, a new dictionary was obtained with overall 85.2% correctness. Moreover, the word pairs in the output dictionary were separated into 11 groups in accordance with the iteration order and were evaluated independently. The evaluation revealed that about

32,000 word pairs were extracted at the first iteration, which take up roughly 50% of the total pairs, with a correctness of 95.3%. In short, the evaluation result showed that we can perform, using this framework, automated creation of a highly accurate bilingual dictionary.

Chapter 4 presents a constraint approach to pivot-based bilingual dictionary induction.
Using a third language to link two other languages is a well-known solution, and usually requires only two input bilingual dictionaries A-B and B-C to automatically induce the new dictionary, A-C. This approach has, however, never been demonstrated to utilize the complete structure of the input bilingual dictionaries, and this is a key problem because the data incompleteness negatively influences the induction result. This chapter describes a constraint optimization-based solution which enhances the quality of pivot-based bilingual dictionary induction. In other words, in this proposal, the lexicon similarity of intra-family languages are realized as semantic constraints and the structure of the input dictionaries are modeled as a Boolean optimization problem based on these constraints which is then formulated within the Weighted Partial Max-SAT (WPMax-SAT) framework, an extension of Boolean satisfiability. All of the encoded CNF (Conjunctive Normal Form), the predominant input language of modern SAT/MAX-SAT solvers, and formulas are evaluated by a solver to produce an optimally correct bilingual dictionary. Moreover, an alternative formalization within the 0-1 Integer Linear Programming framework (also known as Pseudo-Boolean Optimization) was discussed as a comparison study regarding the computational complexity. A tool was designed as an implementation of the proposal using Sat4j, an open source SAT solving library, and IBM CPLEX, a widely used solver of Integer Linear Programming. Using this tool, the proposal was evaluated by inducing an Uyghur-Kazakh bilingual dictionary from Chinese-Uyghur and Chinese-Kazakh dictionaries. As a result, the new dictionary gained 83.7% precision, which is about 10% higher than a baseline method, and 79.5% recall.

Chapter 5 describes an extension to the constraint approach to the pivot-based bilingual dictionary induction.
In the process of the pivot-based bilingual dictionary induction, an additional input dictionary may provide additional information for measuring the semantic relativeness of the cross-lingual word pairs, which is key to suppressing the wrong sense matches. This is because the incompleteness of the existing dictionaries varies and it is reasonable to use the complete part of each dictionary as a shared information for measuring the semantic relativeness. Taking this into account the chapter details a proposal of an extended constraint approach to creating bilingual dictionaries of intra-family languages from more than two input dictionaries. As for the formulization, 0-1 Integer Linear Programming framework was preferred because the dramatic increase in the size of cardinality constraints due to an additional input dictionary is hardly handled by WPMax-SAT due to its dependence on poorly propositional logic. For an evaluation purpose, new dictionaries of Uyghur, Kazakh and Kyrgyz languages were induced from their dictionaries with Chinese, where Kyrgyz is also a member of Turkic language family. The inductions using two and three input dictionaries were conducted, respectively, to

observe the effect of an additional input dictionary on the induction quality. As a result, although the degree of the improvement varies from one language pair to another, an improvement was achieved for all language pairs when the three dictionaries were used as an input. On average, 4%, 2.6% and 4% gains in precision, recall and F-measure were achieved, respectively, which show the effect of the proposal of utilizing more existing bilingual dictionary resources.

Chapter 6 provides the highlights of a bilingual dictionary induction software which was designed as an implementation of the proposals in this thesis.

Chapter 7 summarizes the original contributions and future directions. To this end, this chapter discusses two possible extensions to the proposals 1) using human as heuristics and 2) incorporation of the heuristic framework and the constraint approach to the pivot-based bilingual dictionary induction.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３,０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１,１００wordsで作成し
審査結果の要旨は日本語５００～２,０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、低資源言語のための言語資源構築を目的とするものである。そのために、ピボット（中心軸となる言語を意味する）を用いる技術に基づいて、同族言語間の対訳辞書を自動生成する手法を提案している。得られた主要な成果は以下の通りである。

1. ヒューリスティックスを用いたピボットに基づく対訳辞書生成

低資源言語の対訳辞書を自動生成するには、限られた言語資源を最大限に利用することが重要となる。そこで、言語資源が比較的豊富な言語をピボットとして、低資源の同族言語間の辞書を自動生成する枠組を考案している。まず、異なる言語の単語対の意味的関連度を計算する基本的なヒューリスティックス群を提案している。次に、このヒューリスティックス群を用いて、既存のウイグル語-中国語間の対訳辞書とカザフ語-中国語間の対訳辞書から、新たにウイグル語-カザフ語間の対訳辞書を自動生成する実験を行っている。その結果、提案方式により、高い精度を持つ辞書が生成可能であることを示している。

2. ピボットに基づく対訳辞書生成のため制約最適化アルゴリズム

ピボットを用いて二つの言語をつなぐことは良く知られた手法であり、一般に言語A-B間の対訳辞書と言語B-C間の対訳辞書を用いて、言語A-C間の対訳辞書を自動生成することができる。しかし従来のアプローチでは、入力辞書のデータ量が十分でないと、異なる言語における単語間の意味的関連度を十分に捉えきれず、高い精度の辞書を得ることは困難であった。そこで、本論文では制約最適化アルゴリズムを用いて単語間の意味的関連度を推定する手法を提案している。即ち、同族言語間の語義の類似性を意味的制約として表現し、対訳辞書の生成をこれらの制約に基づく最適化問題としてモデル化している。ウイグル語-カザフ語間の対訳辞書を、ウイグル語-中国語間および中国語-カザフ語間の対訳辞書から生成した結果、従来手法に比べて高い精度の対訳辞書が得られることを示している。

3. 対訳辞書生成の高精度化のための入力対訳辞書数の拡大

入力となる対訳辞書を追加することが、単語間の意味的関連度の推定をより正確にするとの知見に基づき、ピボットに基づく対訳辞書生成のため制約最適化アルゴリズムの拡張を行った。制約数が増大するため解法には0-1線形計画問題を用いている。ウイグル語、カザフ語、キルギス語間の辞書を、各言語と中国語との対訳辞書を入力として生成する実験を行っている。対訳辞書の追加が精度に与える影響を測定するため、入力となる対訳辞書が2個の場合と3個の場合を比較した。その結果、全ての言語対に対して3個の辞書を入力とした場合に生成される対訳辞書の精度が高く、本提案が言語資源の有効活用に資することを示している。

本研究の成果物として、実際に、ウイグル語-カザフ語（5万語）、ウイグル語-キルギス

語（2万5千語）、カザフ語-キルギス語（2万5千語）の対訳辞書が得られている。これらの言語資源は、言語資源のプラットフォームであるLanguage GridやLREMAPで公開されており、テュルク語系言語間の機械翻訳システムに利用されている。

以上、本論文は(1)ヒューリスティックスを用いたピボットに基づく対訳辞書生成、(2)ピボットに基づく対訳辞書生成のための制約最適化アルゴリズム、(3)対訳辞書生成の高精度化のための入力対訳辞書数の拡大を提案し、低資源言語間の対訳辞書が自動生成可能であることを示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成27年2月19日に実施した論文内容とそれに関連した試問の結果、合格と認めた。

注）論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降